"Improving post-editing and automatic translation by the creation of phraseological databases: an experiment"

Colson, Jean-Pierre

Abstract

In spite of the success of phraseology across a range of linguistic disciplines such as corpus linguistics, discourse analysis or semantics, it may come as a surprise that the notion is hardly mentioned in Translation Studies. Delisle (2003), for instance, treats set phrases as part of the lexicon. They are also most conspicuously absent from the major reference work in the field, the Routledge Encyclopedia of Translation Studies (Baker and Saldanha 2011). The same holds true of collocations. As a matter of fact, the interest for phraseology in translation studies came mainly from the European Society for Phraseology (Europhras) and from corpus linguistics (e.g. Teubert 2002). Computational linguistics, in its turn, is showing a growing interest for matters involving translation and collocations in the broad sense. It is now generally recognised that phraseology poses a serious problem to machine translation (MT), because it involves a higher semantic level that cannot be grasped by ...

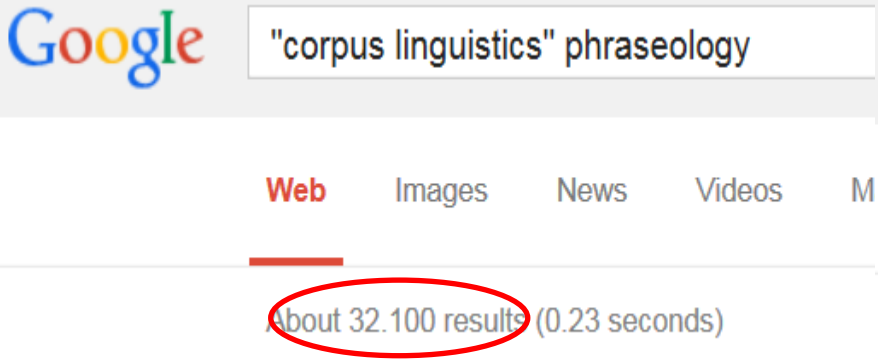
Document type : *Communication à un colloque (Conference Paper)*

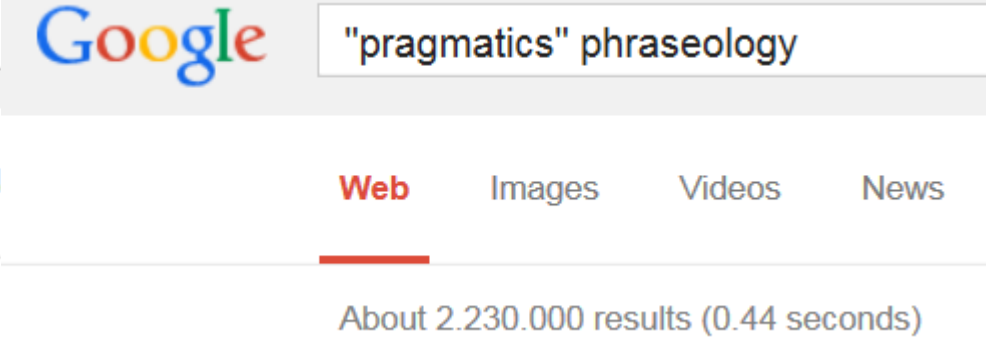
Improving post-editing and automatic translation by the creation of phraseological databases: an experiment

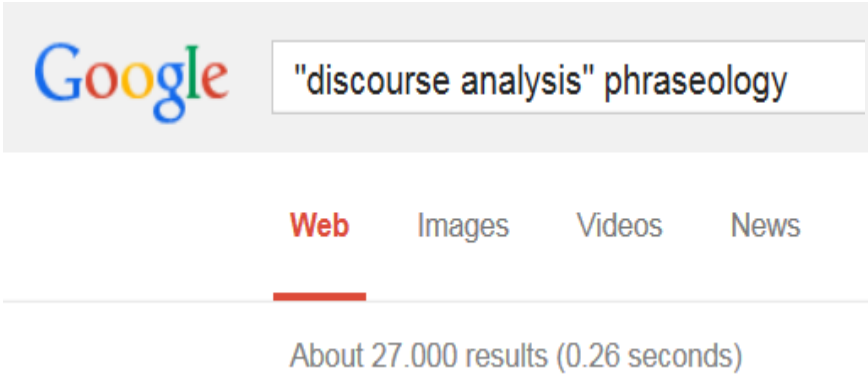
Jean-Pierre Colson
Université catholique de Louvain (Louvain-la-Neuve)

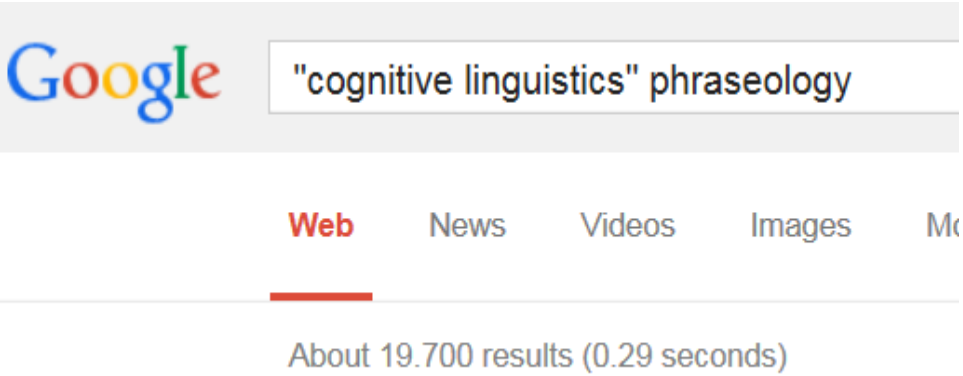
1. Is Phraseology ignored by Translation Studies?

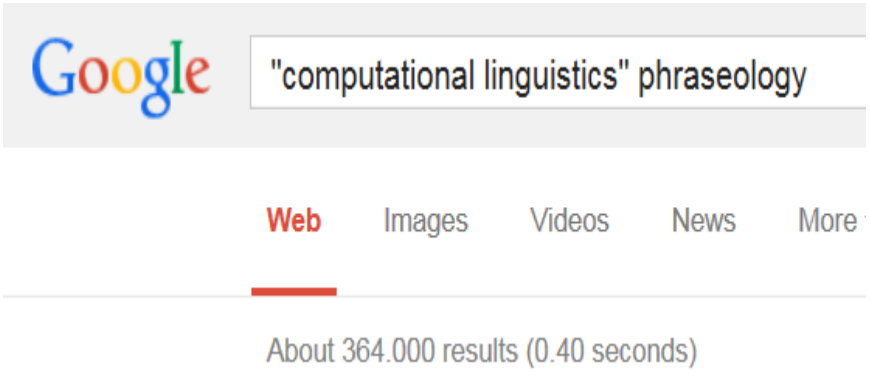
* We refer here to phraseology as the study of set phrases in the broadest sense, including partly fixed phrases (routines and formulae, collocations), and also very fixed phrases (idioms and proverbs).
* The notion of *phraseology* is now used across a wide range of linguistic disciplines: Phraseology (proper), Corpus Linguistics, Discourse Analysis, Pragmatics, Cognitive Linguistics, Computational Linguistics (…)

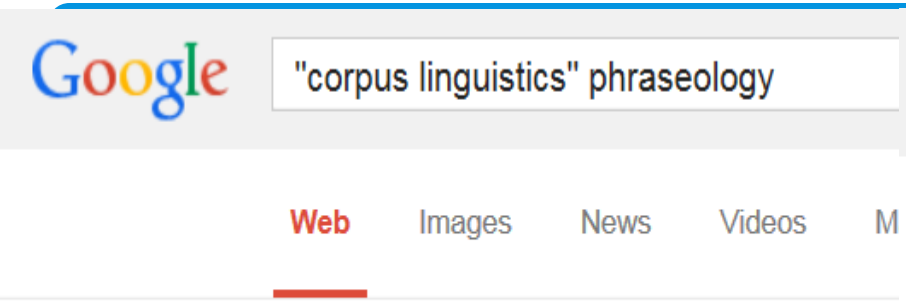

Google
"corpus linguistics" phraseology
Web
Images
News
Videos
M
About 32.100 results (0.23 seconds)

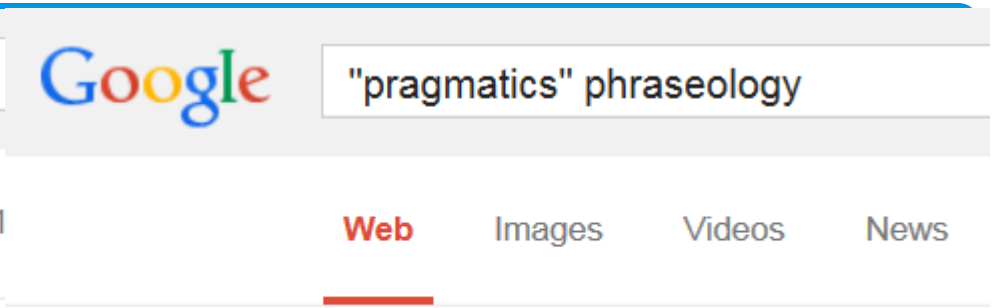

Google
"pragmatics" phraseology
Web
Images
Videos
News
About 2.230.000 results (0.44 seconds)

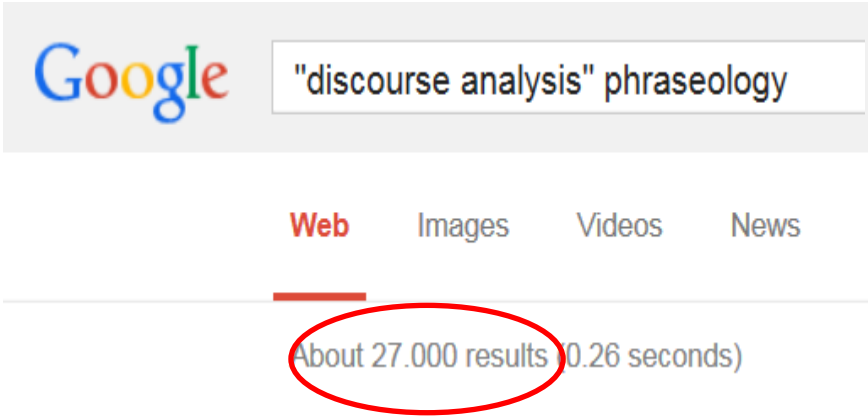

Google
"discourse analysis" phraseology
Web
Images
Videos
News
About 27.000 results (0.26 seconds)

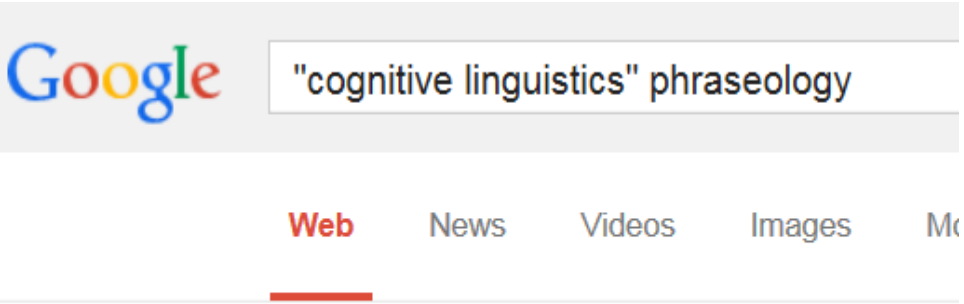

Google
"cognitive linguistics" phraseology
Web
News
Videos
Images
About 19.700 results (0.29 seconds)

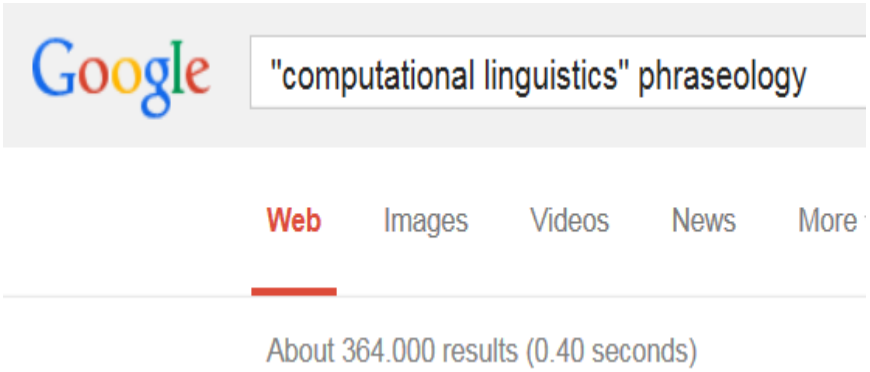

Google
"computational linguistics" phraseology
Web
Images
Videos
News
More
About 364.000 results (0.40 seconds)

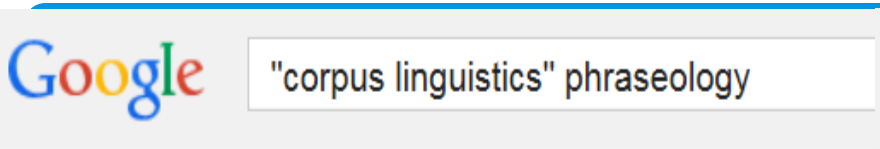

Google
"corpus linguistics" phraseology
Web
Images
News
Videos
M
About 32.100 results (0.23 seconds)

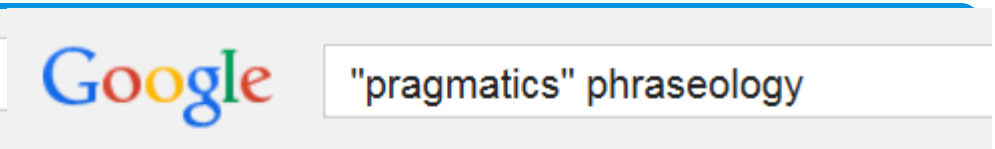

Google
"pragmatics" phraseology
Web
Images
Videos
News
About 2.230.000 results (0.44 seconds)

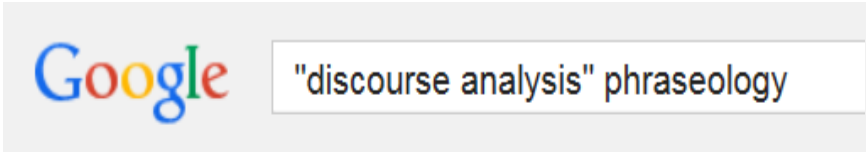

Google
"discourse analysis" phraseology
Web
Images
Videos
News
About 27.000 results 0.26 seconds)

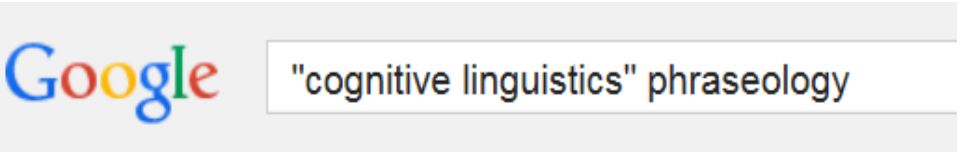

Google
"cognitive linguistics" phraseology
Web
News
Videos
Images
About 19.700 results (0.29 seconds)

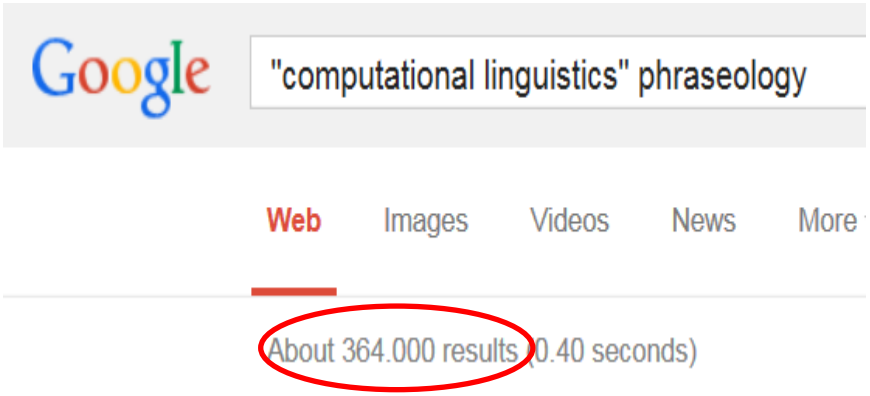

Google
"computational linguistics" phraseology
Web
Images
Videos
News
More
About 364.000 results (0.40 seconds)

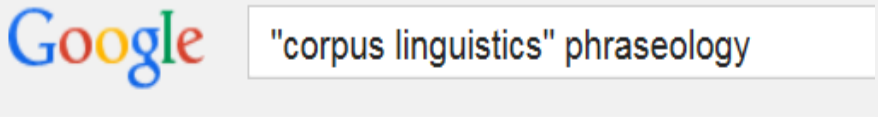

Google
"corpus linguistics" phraseology
Web
Images
News
Videos
M
About 32.100 results (0.23 seconds)

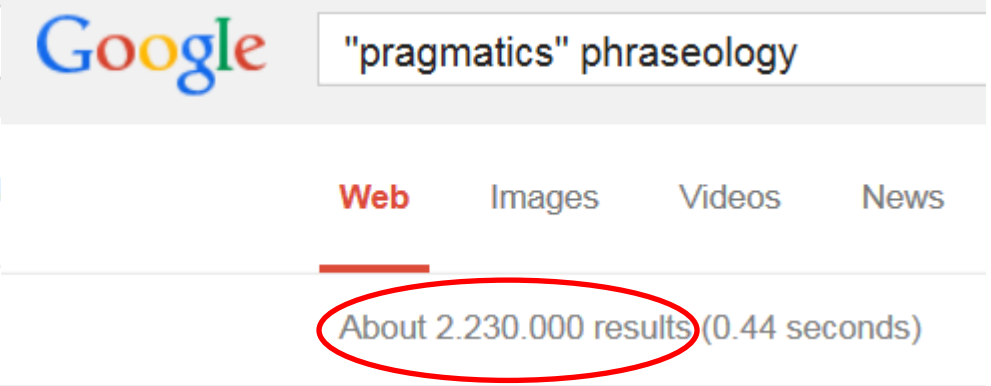

Google
"pragmatics" phraseology
Web
Images
Videos
News
About 2.230.000 results (0.44 seconds)

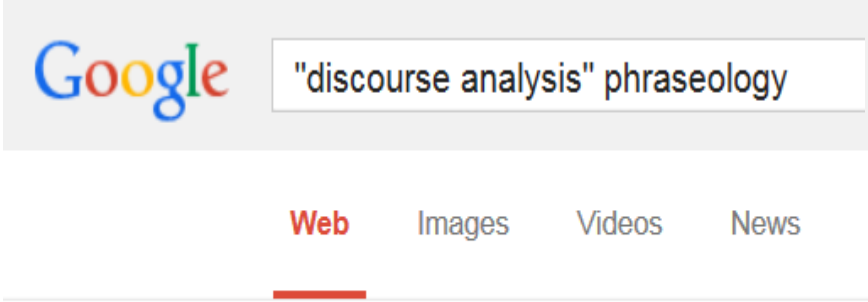

Google
"discourse analysis" phraseology
Web
Images
Videos
News
About 27.000 results (0.26 seconds)

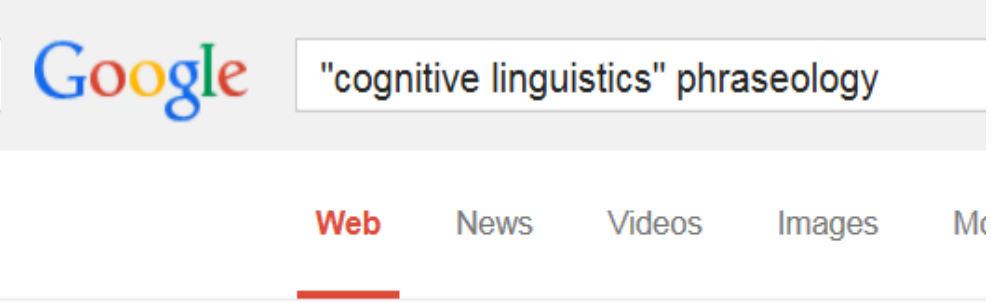

Google
"cognitive linguistics" phraseology
Web
News
Videos
Images
About 19.700 results (0.29 seconds)

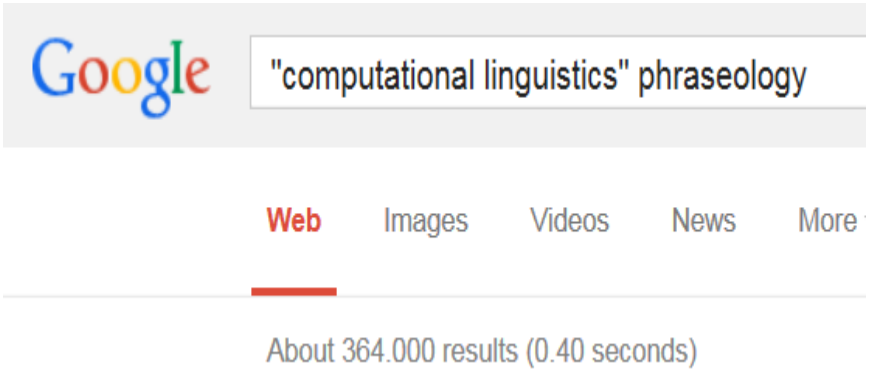

Google
"computational linguistics" phraseology
Web
Images
Videos
News
More
About 364.000 results (0.40 seconds)

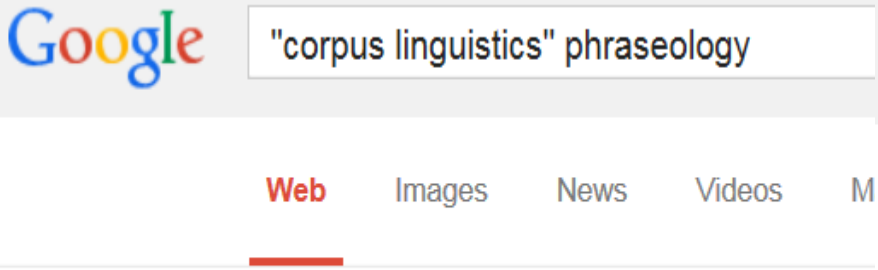

Google
"corpus linguistics" phraseology
Web
Images
News
Videos
M
About 32.100 results (0.23 seconds)

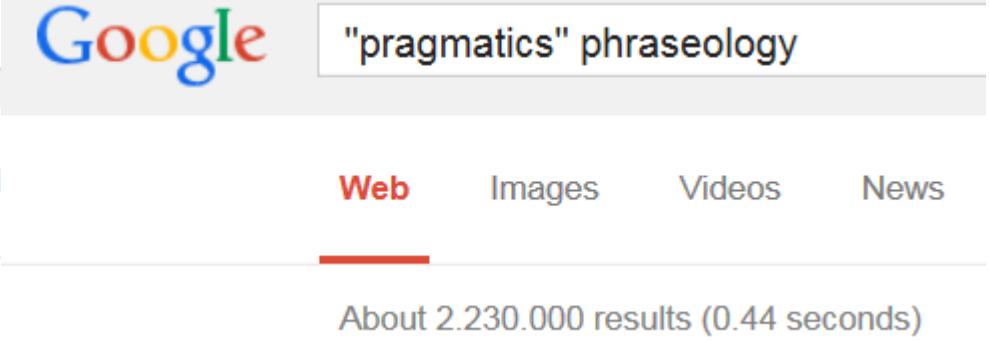

Google
"pragmatics" phraseology
Web
Images
Videos
News
About 2.230.000 results (0.44 seconds)

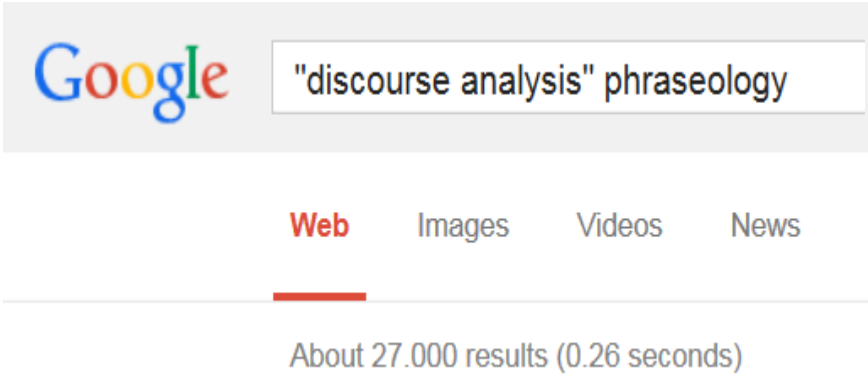

Google
"discourse analysis" phraseology
Web
Images
Videos
News
About 27.000 results (0.26 seconds)

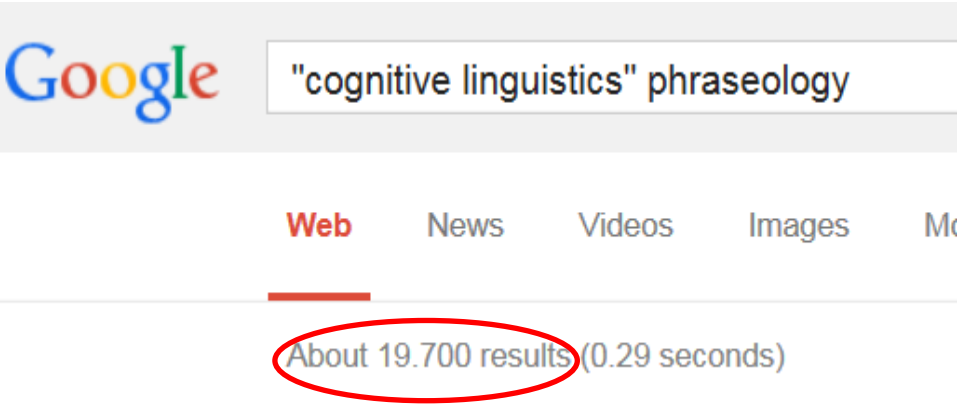

Google
"cognitive linguistics" phraseology
Web
News
Videos
Images
About 19.700 results (0.29 seconds)

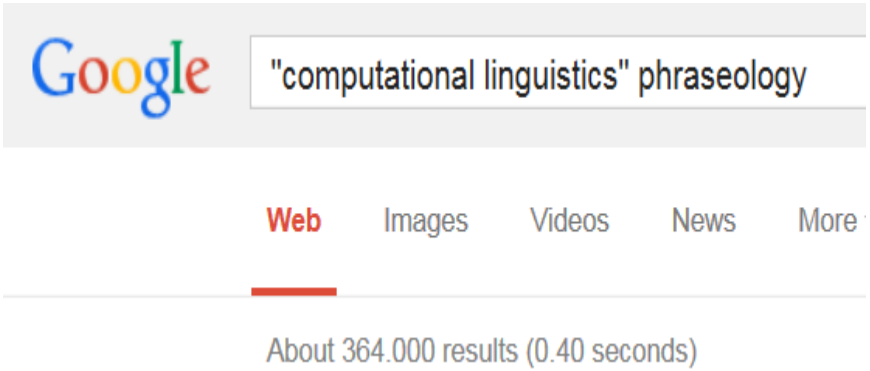

Google
"computational linguistics" phraseology
Web
Images
Videos
News
More
About 364.000 results (0.40 seconds)


Google
"corpus linguistics" phraseology
Web
Images
News
Videos
M
About 32.100 results (0.23 seconds)


Google
"pragmatics" phraseology
Web
Images
Videos
News
About 2.230.000 results (0.44 seconds)


Google
"discourse analysis" phraseology
Web
Images
Videos
News
About 27.000 results (0.26 seconds)


Google
"cognitive linguistics" phraseology
Web
News
Videos
Images
About 19.700 results (0.29 seconds)


Google
"computational linguistics" phraseology
Web
Images
Videos
News
More
About 364.000 results (0.40 seconds)

* What about Translation Studies?
* Most publications on Translation Studies mention the problem of *expressions / idioms / collocations but they do not refer to phraseology* as a theory or discipline
* Example 1: Delisle, J. (2003). *La traduction raisonnée*: *expressions* are treated as a part of the lexicon

* Example 2: phraseology is also most conspicuously absent from a major reference work in the field, the *Routledge Encyclopedia of Translation Studies* (Baker and Saldanha 2011), and the same holds true of *collocations.*

* The interest for phraseology (at least for collocations) in translation studies came mainly from corpus linguistics (example: W. Teubert, 2002. The Role of Parallel Corpora in Translation and Multilingual Lexicography).

* On the other hand, *computational linguistics* is now showing a growing interest for phraseology, particularly against the backdrop of automatic translation
* Monti, Mitkov, Corpas Pastor, Seretan, eds. (2013), Workshop Proceedings: Multi-word units in machine translation and translation technologies, Nice Machine Translation Summit.

* The notion of *multiword unit* comes close to that of phraseology in the broad sense, but it also includes *lexical bundles* (Doug Biber), n-grams selected on the sole basis of their recurrent frequency (20 to 40 PMW, cfr. Biber, Conrad et Cortes 2004)

2. Can automatic translation really cope with phraseology?

* Several studies have clearly demonstrated that phraseology in the broad sense remains problematic for automatic or semi-automatic translation:
* Sag, I.A. et al (2001). Multiword Expressions: A Pain in the Neck for NLP

* Rayson, P. et al (2010). Multiword expressions: hard going or plain sailing?
* Barreiro, A. et al. (2013). When Multiwords Go Bad in Machine Translation
* The general picture is that even the best systems (*Google Translate, OpenLogos, Systran*, etc.) produce erroneous translations in about 50 percent of the cases

* Example: results from Barreiro et al. (2013) with *OpenLogos* and *Google Translate* (language combinations: English, French, Italian, Portuguese)

System	Lang pair	OK	ERR	Total
OL	EN-FR	40	48	88
	EN-IT	36	83	119
	EN-PT	60	96	156
	Total	136	227	363
GT	EN-FR	70	38	108
	EN-IT	59	47	106
	EN-PT	67	47	114
	Total	196	132	328

* Recent examples from *Google Translate*


Google
Traduction
Allemand
Français
Anglais
Tilni aniqlash
Anglais
Français
Allemand
Traduire
Drip, drip, drip; and the next thing you know, democracy has been washed away
Goutte à goutte, goutte à goutte, goutte à goutte; et la prochaine chose que vous savez, la démocratie a été emporté


SYSTRANet
Your free online translator


Text
Web Page
RSS
File
Dictionary
My Dictionary
Register for
From:
English
To:
French
Translate
Options:
Tim Cahill happy to sign off from World Cup stage with a bolt from the blue
Tim Cahill heureux de se déconnecter de l'étape de coupe du monde avec un boulon du bleu

Anglais | Allemand | Français | Tilni aniqlash

Français | Allemand | Anglais | Traduire

La réforme de l'Etat, l'Arlésienne du quinquennat
LE MONDE | 28.06.2014 à 10h59 • Mis à jour le 28.06.2014 à 10h59 |
Par Bertrand Bissuel
Abonnez-vous
à partir de 1 € Réagir ClasserPartager facebook twitter google + linkedin pinterest

Thierry Mandon, avec Manuel Valls et Najat Vallaud-Belkacem, à l'Elysée, après le conseil des ministres, le 18 juin.
Passer à la vitesse supérieure. C'est dans cet état d'esprit que Thierry Mandon accomplit ses premiers pas de secrétaire d'Etat à la réforme de l'Etat et à la simplification. Nommé à ce poste le 3 juin, l'ex-porte-parole du groupe PS à l'Assemblée nationale hérite d'un dossier – la modernisation de l'action publique (MAP) – qui n'a pas produit les résultats espérés.

The reform of the state's five-year Arlesienne
THE WORLD | 28/06/2014 at 10:59 • Updated 28/6/2014 at 10:59 |
Bertrand Bissuel
Subscribe
from € 1 Comment ClasserPartager facebook twitter google + linkedin pinterest

Thierry Mandon, Manuel Valls and Najat Vallaud-Belkacem, at the Elysee Palace after the cabinet meeting on June 18
Move up a gear. It is in this spirit that Thierry Mandon takes his first steps as Secretary of State for State reform and simplification. Appointed to this position on June 3, the former spokesman for the PS group in the National Assembly inherits a record - modernization of public policy (MAP) - which did not produce the expected results.

* The underlying causes of these shortcomings have been clearly identified by the previously mentioned studies
* The main reason is the insufficient inventory of all set phrases in the broad sense by lexicographic reference works, even for major languages such as English
* An essential first step in order to improve the quality of automatic translation, and also to facilitate post-editing, is therefore improving the efficiency of the automatic extraction of phraseology

* In our experiment, we have tried to start from a web corpus of reasonable size (200 MW), a portion of the ukWacky corpus (Baroni et al. 2009)
* All n-grams (from size 2 to 9) were extracted and filtered by a combination of frequency (>2) and a new statistical score: the *Corpus Proximity Ratio* (*CPR*, J.-P. Colson 2014), measuring the degree of attraction between the component parts of an n-gram

$$cpr = \frac{n\,(x_{i_1} x_{i_2} x_{i_3} \ldots x_{i_n})}{n(\max\limits_{x_{i_1}, x_{i_2}, x_{i_3}, \ldots x_{i_n}} (||x_i - x_j||) \leq W}$$

The Corpus Proximity Ratio (CPR), J.-P. Colson 2014

* Examples : *take the road, hit the road, fork in the road (PerlPr)*
* A database of about 400,000 English candidate collocations has been assembled that way; it lies on the border between low-frequency lexical bundles and phraseology
* Examples: a bit, art, back to, bank, barrel, bridge, bring, jet, road

* *CPR* meets four criteria recommended by Gries (2013) for the improvement of automatic extraction of collocations:
  * *The measure is directional*
  * *The methodology uses recurrence across corpora*
  * *It is extendable (extended) to multiword expressions*
  * *There may be a pyscholinguistic foundation in the Firthian principle of attraction, and in the comparison with association databases*

* Apart from collocations and different types of set phrases, the method makes it possible to extract all kinds of semantic associations between words; for instance, in the case of associations with the English word *art*: *archaeology and art, art and antiques, art and architecture, art and craft, art and culture, art and design, art and science, art at its very best*, etc.

* Although more results are necessary, this also seems to corroborate the psycholinguistic grounding of this methodology: it not only produces collocations and idioms, but also common associations derived from society and culture.
* As also pointed out by Gries (2013), a crucial step in the validation of automatic extraction algorithms is the comparison with reliable psycholinguistic association databases, such as those of the University of South Florida Association Norms (http://w3.usf.edu/FreeAssociation).

* In the case of *art*, the list of overlapping collocations (about 40 percent) includes (in the same order as in the Association norms): *crafts, museum, craft, design, contemporary, science, ceramics, canvas, drama, modern, culture, critic, skill, ceramic, original, performance, architecture, project*
* This raises another crucial question for IR and NLP: to what extent are psycholinguistic associations interwoven with phraseology?

* From a theoretical point of view, the automatic extraction of phraseology poses the question of the statistical nature of language, a crucial issue in statistical machine translation
* According to the Zipfian law (Zipf 1949), the general distribution of words displays a very limited number of high frequency items, a fair amount of average frequency words, and a *long tail* of words with extremely low frequency (an *army of dwarves*)

* Although the matter is still controversial, Zipf himself interpreted his law as evidence for the *principle of least effort* in language
* This principle is closely related to the *principle of economy*: language re-uses the same elements in order to avoid linguistic inflation (Martinet's *double articulation* at the level of morphemes and phonemes, polysemy, phraseology)

* In 1953, the mathematician Mandelbrot proposed a slight correction to Zipf's law; the law of Zipf-Mandelbrot is now the most widely used version

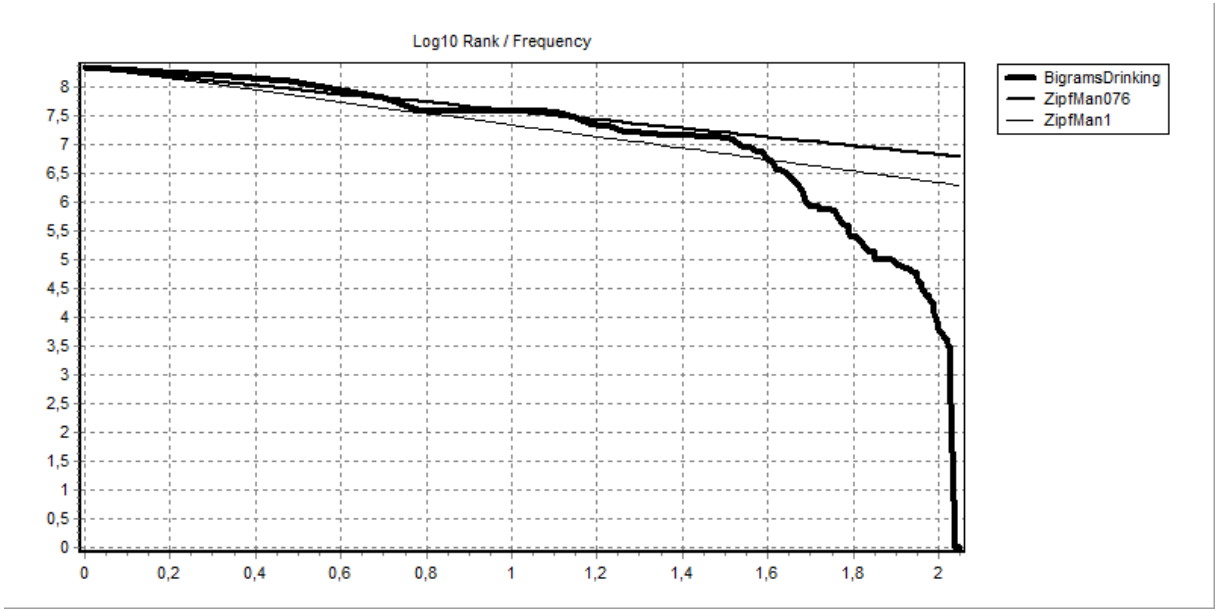
$$f(w) = \frac{C}{(r(w)+b)^a}$$

* where *f (w)* represents the frequency of a word, and *r(w)* its frequency rank. *C* and *a* are constants that are set empirically according to the data. *C* is normally set to the highest frequency value obtained and *a* has been set at 1.09 for the British National Corpus. As for *b*, Baroni (2008) recommends an empirical adaptation by increasing it according to the results, with a typical increase of *b=1* for the 20 highest frequency ranks

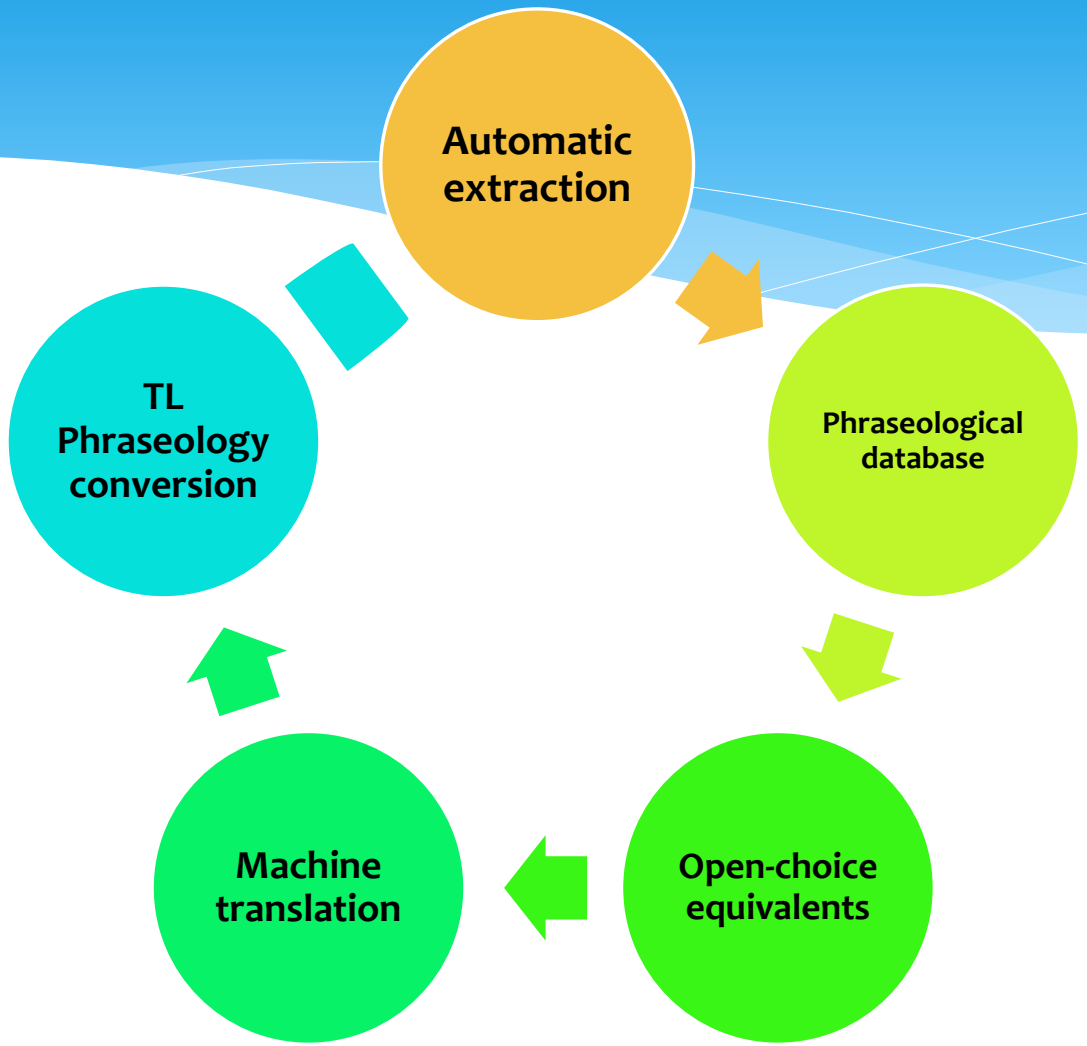
* Mandelbrot's interpretation of Zipf's law was precisely that word combinations in language follow statistical principles; to put it differently, the basis for phraseology is (largely) statistical…
* If this is the case, n-gram frequencies should also display a Zipfian distribution
* This is precisely suggested by Baroni (2008)

* Experiments tend to show that this is indeed the case, even at the level of small texts
* Example:
* “I normally drink between 18 and 24 units per week. I consider the week worryingly quiet if I haven’t had cause to drink one bottle of wine by Wednesday. Tuesday feels far enough into the week to be a heavy night, then the next night is usually a day of relative rest. If I don’t have another big night on Thursday, I will let loose on Friday or Saturday. That’ll be a good bottle. Sunday is generally a recovery night. I pondered the dry week ahead. I realised I had not had one in four years. I felt like a voyager setting off to a featureless land. At 23, did I really need alcohol to enjoy my evenings?” (The Times Online, August 03, 2005)

* x stands for the log10 of the rank of the items (shown in decreasing order of frequency) and y for the log of the frequency for each item. If the Zipf-Mandelbrot principle applies here, such a log-log table should (for mathematical reasons) display a straight line, with an abrupt fall at the right end of the table and some minor irregularities along the line (Mandelbrot's corrections).
* The next fiure presents the log-log results for the bigrams (*BigramsDrinking*), as well as two projections according to the law of Zipf-Mandelbrot. We follow here the same method as Ha/Sicilia-Garcia/Ming/et al. (2002), who have computed bigram frequencies for the whole Brown Corpus (1,000,000 words): no *b* factor is used, and *a* is set to 0.76 (in the legend of Figure 1: *ZipfMan076*). By way of comparison, Figure 1 also presents a projection with *a=1* (*ZipfMan1).*


Log10 Rank / Frequency
BigramsDrinking
ZipfMan076
ZipfMan1
8
7,5
7
6,5
6
5,5
5
4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
0
0,2
0,4
0,6
0,8
1
1,2
1,4
1,6
1,8
2

* From the point of view of translation, a database of (even partial) phraseological n-grams can be useful for extracting potentially problematic structures
* It could also find a place in a *simplified semantic* model


Automatic extraction
Phraseological database
Open-choice equivalents
Machine translation
TL Phraseology conversion